Early Identification of At-Risk Students Using Iterative Logistic Regression

Li Zhang and Huzefa Rangwala

Department of Computer Science,
George Mason University, Fairfax, Virginia, USA
lzhang18@gmu.edu, rangwala@cs.gmu.edu

Abstract. Higher education institutions are faced with the challenge of low student retention rates and high number of dropouts. 41% of college students in United States do not finish their undergraduate degree program in six years, and 60% of them drop out in their first two years of study. It is crucial for universities and colleges to develop data-driven artificial intelligence systems to identify students at-risk as early as possible and provide timely guidance and support for them. However, most of the current classification approaches on early dropout prediction are unable to utilize all the information from historical data from previous cohorts to predict dropouts of current students in a few semesters. In this paper, we develop an Iterative Logistic Regression (ILR) method to address the challenge of early prediction. The proposed framework is able to make full use of historical student record and effectively predict students at-risk of failing or dropping out in future semesters. Empirical results evaluated on a real-wold dataset show significant improvement with respect to the performance metrics in comparison to other existing methods. The application enabled by this proposed method provide additional support to students who are at risk of dropping out of college.

Keywords: Iterative Logistic Regression, educational data mining, early dropout prediction

1 Introduction

According to the National Center for Education Statistics [17], more than 41% of students who began seeking an undergraduate degree at a four-year institution in Fall 2009 failed to graduate within six years. In 2008–2009, higher education institutions spent more than \$263 billion on education and related expenses, delivering the equivalent of \$487.5 million semester-credit hours of instruction. The spent amount has grown up to \$536 billion in academic year 2014-2015. According to recent work by Schneider [20], US taxpayers spent more than \$9 billion providing education to first year students who failed to return the following year. Schneider and Yin [21] further estimated the opportunity cost for college dropouts from just a single cohort of entering students lost is \$3.8 billion in lifetime income, and the government loss is at \$730 million in potential tax revenue.

Prior research work in this area has involved the analysis of student dropout data and identified the need for early determination and possible intervention and additional support from the institution [22]. Dynarski [9] provides several possible practical intervention ideas for secondary school dropouts which can be implemented at the university level. There is also evidence that summer bridge programs are helpful in learning and producing environments that can ultimately improve student performance and retention [5, 16].

There are many different definitions that have been used for the term 'dropout' and 'retention' based on the count of returning next-year students and graduating in 6 years. Typically retention is staying at school until completion of a degree and dropping out is leaving school permanently without a degree. However, a student might leave the college to work a few years and come back to finish the degree. This is known as *stopout* [14]. More than three decades ago, Alexander Astin identified the dropout as a problematic concept [3]. According to Astin, "a 'perfect' classification of dropouts versus non-dropouts could be achieved only when all of the students had either died without ever finishing college or had finished college." Hagedorn [13] discussed this vague and complicated definition of retention/persistence and concluded there is not a consensus on the correct way to measure retention – it depends on the context. To accomodate both dropout and stopout, our study use a stricter definition: a *dropout student* is defined as a student who fails to register the next semester or has a GPA of 0.0 in the next semester. Our goal is to effectively predict and identify students who are at a high risk of dropping out. We propose an Iterative Logistic Regression (ILR) to predict student dropouts, which has interpretable coefficients and identifies at-risk students at an early stage.

Our paper has several main contributions. We propose a robust method with regularization to model the sequential effect of previous term performance. We are able to learn from all the semesters from historical data of the past cohorts of students and effectively generate important features for dropout prediction in future semesters. Our proposed model has the further advantage of easily interpreted predictions.

2 Literature review

In recent years, research work has explored key features using classical statistical methods to identify students at risk of dropping out from their field of study and leaving college/university. Golding et al.. [12] identified the relationship between students' overall academic performance (GPA) and matriculation chances in the first year based on enrollment information. Druzdzel and Glymour [8] were among the first researchers to apply machine learning algorithms to study the student retention problem. Campbell [6] used factor analysis and logistic regression on a set of student features derived from data extracted from Blackboard [4]. Pittman [19] compared several data mining techniques (logistic regression, decision tree, Bayesian classifiers and neural network) to predict student retention and concluded that logistic regression had the best performance. Logistic regression has also been used in different contexts for early dropout prediction [23, 11]. Kovacic [15] explored the effect of demographic variables and study environment on the outcome of student enrollment with various tree-based methods and logistic regression. Nandeshwar et al. [18] analyzed retention records and concluded that focusing more resources on high-risk groups of students is helpful in improving their chances of completing a university degree. Baradwaj and Pal [4] applied decision tree to classification using features derived from attendance reports, class test scores and assignment submissions. Tanner [24] used the k-nearest neighbor method to predict student failure in an online course setting. Although these methods are able to predict student dropout, none of them are able to make full use of the extra semesters in the data from the previous student cohort for early dropout prediction. Ameri et al. [2] and Chen et al. [7] performed survival analysis, particularly cox regression, on the student performance datasets. Their analysis does not take into account the additional correlation in the data after the first dropout occurs and hence is not able to model student stopout. Our proposed method is able to work with the extra available semesters in the past student record and does not assume one-time dropout.

3 Methods

The primary objective of this study is to predict student dropouts in future semesters. A *dropout student* is defined as a student who fails to register a semester or has a zero GPA in the semester. Table 1 is a summary of notations which we use in this paper. Let n be the number of semesters a student is typically at a university. We have p_t students at the beginning of each semester, where p_t is different in each semester. Let $\boldsymbol{R}$ denote time invariant features of students (such as demographic information). We use $\boldsymbol{G}_t = [\boldsymbol{g}_1, \boldsymbol{g}_2, \cdots, \boldsymbol{g}_t]$ to represent semester GPA from 1 to t-th term. We also developed additional features including *Absence* denoted by $\boldsymbol{A}_t$; generated by GPA in t-th semester. $\boldsymbol{A}_t = [\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_t]$ represents absence and is indicated by semester GPA $\boldsymbol{g}_t$. For Student j in t-th term,

$$a_{tj} = \begin{cases} 0 & g_{tj} \geq 0 \\ 1 & \text{Otherwise} \end{cases} \tag{1}$$

We use y_{tj} as labels for a dropout event in t-th term for student j. $y_{tj} = 1$ implies that student j is considered a dropout.

Logistic regression is a technique that has been widely used by researchers in the field of education data mining. Observing student performance in previous semesters usually has an impact on his future performance, we focus on using this fact to improve logistic regression.

Logistic regression Let m be the number of semesters that we have GPA for in the test set. $\boldsymbol{x}_{tj}$ denotes all features including static and time-dependent

Table 1. Glossary of Symbols

Notation	Description
s	number of students by semester t
m	number of available semesters of new students available
n	number of semesters of new student we predict up to
$\boldsymbol{x}_t$	all predictor variables for the set of students existing for prediction in semester t
$\boldsymbol{x}_{tj}$	vector of predictor variables for Student j existing for prediction in semester t
$\boldsymbol{G}_t$	matrix containing GPA record up to semester t
$\boldsymbol{g}_{tj}$	GPA record for semester t for student j
$\boldsymbol{A}_t$	matrix containing feature of absence up to semester t
$\boldsymbol{a}_{tj}$	record of absence for semester t for student j
$\boldsymbol{R}$	matrix containing static features such as demographic feature
$\boldsymbol{\omega}_t$	the weight vector for semester t
$\boldsymbol{\lambda}$	L_1 regularization parameter for over-fitting
$\boldsymbol{y}_t$	set of labels for students by semester t
y_{tj}	the label for student j in semester t
$\boldsymbol{p}_t$	predicted probability for dropout for semester t

features for student j in the t-th semester. d_{tj} is the length of the feature vector $\boldsymbol{x}_{tj}$.

y_{tj} are class labels representing the event of dropout for student j in t-th semester (1 for dropout and 0 otherwise). Logistic regression computes the probability that student j dropouts in a semester by:

$$\begin{aligned} p(y_{tj}|\boldsymbol{x}_t;\boldsymbol{\omega}_t) &= \sigma(y_{tj}\boldsymbol{\omega}_t^{\mathsf{T}}\boldsymbol{x}_{tj}) \\ &= \frac{1}{1+\exp\left(-y_{tj}\boldsymbol{\omega}_{tj}^{\mathsf{T}}\boldsymbol{x}_{tj}\right)} \end{aligned} \tag{2}$$

where $\boldsymbol{\omega}_t = [\omega_{i1}, w_{i2}, \cdots, \omega_{id_t}]^{\mathsf{T}}$ is the coefficient vector to be solved for .

We impose the L_1 constraint [25] to enforce sparsity. This estimator is known as lasso and its L_1-regularized log-likelihood is given by:

$$l(\boldsymbol{\omega}_t) = \sum_{j=1}^{s_t} \log(1+\exp(y_{tj}\boldsymbol{\omega}_t^{\mathsf{T}}\boldsymbol{x}_{tj})) + \frac{\lambda}{2}|\boldsymbol{\omega}_t| \tag{3}$$

λ is a parameter that we need to tune. The L_1 penalty is used for both variable selection and shrinkage. A sufficiently large λ will cause some coefficients equal to 0 and hence not be included in the model. By setting appropriate λ value, we are able to eliminate misleading or unnecessary features and make the model easier to interpret. ω_i can be solved by the Proximal Newton method proposed by Friedman et al. [10].

Iterative Logistic Regression The key idea behind the iterative logistic regression is to incorporate the predicted outputs that denote the probability of

Algorithm 1: Algorithm for Iterative Logistic Regression (ILR).

```
Data: $m$, $n$, training data $\boldsymbol{x}_n$, testing $\boldsymbol{x}'_m$
Result: testing data $\boldsymbol{x}'[(m+1):n]$
parameter : $\lambda_L$, $\lambda_U$
Extract $\boldsymbol{G}_n$, $\boldsymbol{A}_n$, $\boldsymbol{R}$ from $\boldsymbol{x}_n$;
Extract $\boldsymbol{G}'_m$, $\boldsymbol{A}'_m$, $\boldsymbol{R}'$ from $\boldsymbol{x}'_m$;
$\boldsymbol{x} = [\boldsymbol{A}_n[1:m], \boldsymbol{G}_n[1:m], \boldsymbol{R}]$;
$\boldsymbol{x}' = [\boldsymbol{A}'_n[1:m], \boldsymbol{G}'_n[1:m], \boldsymbol{R}']$;
for $t = (m+1)$ to $n$ do
    $\boldsymbol{y}_t = \boldsymbol{A}[t]$;
    Learn $\boldsymbol{\omega}_t$ by Eq.(2) and (3) ;
    Estimate $\boldsymbol{p}_t$ into Eq.(2);
    $\boldsymbol{x} = [\boldsymbol{x}, \boldsymbol{p}_t]$ ;
    Substitute $\boldsymbol{\omega}_t$ into Eq.(2) to estimate $\boldsymbol{p}'_t$;
    $\boldsymbol{x}' = [\boldsymbol{x}', \boldsymbol{p}'_t]$;
end
```

dropout in earlier semesters for future semesters in a cascaded manner. Given m semesters of historical data as a first step we use the LR model with lasso penalty to obtain dropout prediction for the $(m+1)$-th semester. Besides the time-invariant features, we use this probabilistic value as an additional variable to include for training a new regularized LR model for identifying dropouts in the $(m+2)$-th semester. Generalizing this further, the estimated probability $\hat{p_k}$ of dropout in a future semester k $(m < k \leq n)$ can be computed by the following equation:

$$\hat{p}(y_{jk}|\boldsymbol{x}_m) = \sigma\big(\sum \boldsymbol{\omega}_k\big(\sum_{t=m+1}^{k-1} \hat{p}(y_{jt}|\boldsymbol{x}_{mj}) + \sum_{t=1}^{m}(\boldsymbol{A}_t + \boldsymbol{G}_t) + \boldsymbol{R}\big)\big) \tag{4}$$

By Equation 3, $p(y_k|\boldsymbol{x}_m)$ is the probability of dropout and m represents the available terms for testing data. A_t and G_t are the feature of absence and GPA of semester t. (i.e. if $G_j = 0$, $A_j = 1$.) R contains the non time-dependent feature including high school GPA, gender, race, school/department while admitted. We can then substitute $\boldsymbol{\omega}_k$ iteratively to estimate the probability of dropout in the next few years.

4 Experimental Protocol

4.1 Dataset

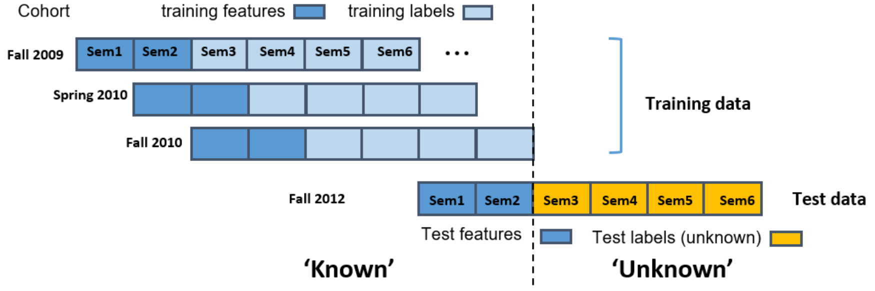
We performed experiments on a dataset from George Mason University (GMU), a large public univeristy in the Unitied States, starting from Fall 2009 to Spring 2016. Since the record of transfer students have less data, we focus on first-time-entry students. The following information from the student grade database are obtained: id, cohort, age, the major they applied, high school GPA, ACT scores, Scholastic Assessment Test (SAT) scores, graduating term, SAT math score, duration, enrollment years and semester GPA.

For evaluation, we only use students who are in cohorts within the Fall 2009 to Spring 2013 ranges. This ensures we have full six semesters for validation. We have a total of 13643 records.

4.2 Experimental Protocol

Fig. 1 shows our evaluation protocol. Assume we are at the end of Spring 2013 and would like to identify dropouts for the students first enrolled in Fall 2012, then there are only data available from Fall 2012 and Spring 2013 for this cohort. We use student enrolled first enrolled from Fall 2009 to Spring 2012 as training set, and the students enrolled from Fall 2012 to Fall 2013 as testing set. There are 7932 students in training set and 5690 students in the testing set. We compare our proposed method with six baseline methods: random guess (RD), Naive Bayes (NB), Random Forest (RF), Decision Tree (DT), k-Nearest-Neighbour (KNN) and Logistic Regression (LR).

Fig. 1. Demonstration for using historical student records from Fall 2009 to Fall 2011 to predict Fall 2012 students dropout in the end of Spring 2013. 'Sem' is short for semester



4.3 Data Pre-processing

Fig. 2 shows the student dropout rate from cohort Fall 2009–2013. The results shows that students enrolled in different years have similar distribution irrespective of their starting semester since they first enrolled. Under this assumption, we are allowed to use the data from the past cohorts of students to predict current students dropouts in the future semesters. We simply align the data by the term they first enrolled in the system. Table 2 and 3 are the sample student data before and after alignment and cleaning. Student A first enrolled in Fall 2015 and obtained a GPA of 3.0 and student B first enrolled in Spring 2016 and obtained a GPA of 3.5. Then student A and student B have a GPA of 3.0 and 3.5 for their first semester, respectively. There are approximately 15-40% of missing data for the high school GPA, ACT scores and SAT score features. The missing values for high school GPA and SAT scores are imputed by the respective mean value. We also take the natural logarithm of SAT scores to avoid scaling issue on regression model.

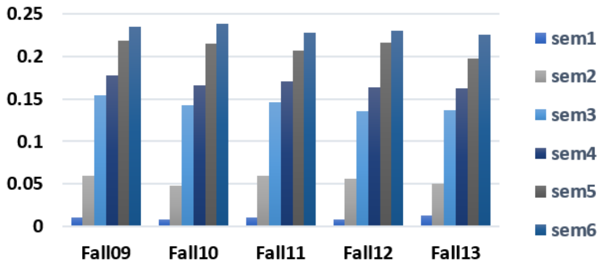


Fig. 2. Barchart on first 6 semesters of student dropout rate for cohort Fall 2009–2013 shows very similar distribution across different cohorts. 'Sem' is short for semester.

Table 2. Sample student data before aligning and cleaning.

id	Cohort	Fall15	Spring16	Fall16	Spring17
1	Fall15	3.0	NA	2	?
2	Spring16	-	2	3.2	?
3	Fall16	-	-	4	?

Table 3. Sample student data after aligning and cleaning. G(t), A(t) denotes GPA and Attendance for semester t. (1 means dropout and 0 means retain).

id	G(1)	G(2)	G(3)	G(4)	A(1)	A(2)	A(3)	A(4)
1	3.0	0	2	0	0	1	0	?
2	2	3.2	1.8	-	0	0	?	?
3	4	4	-	-	0	0	?	?

4.4 Metrics

Our goal is to evaluate the performance of various models on the task of predicting if a student is likely to enroll in the next semester; and hence predict whether the student will drop out or not. The outputs of logistic regression are in the form of probability. Since we are dealing with the prediction of a binary outcome, we apply True Positive (TP), False Negative (FN) and False Positive (FP) to measure the counts. We use 0.5 as the decision boundary. We also use *AUC*, *PRAUC* and *F1* to accommodate the imbalance issue. AUC is expressed as Area Under the Receiver Operating Characteristic (ROC) where the curve is created by plotting the true positive rate (TPR) against the False Positive rate (FPR) under various threshold values. PRAUC (Precision-Recall Curve) is the curve created by plotting Precision against Recall. This curve is more sensitive on positive class in general. Since we are more interested in positive prediction being correct (precision) without missing students at risk (recall), we pick PRAUC as our evaluation metrics.

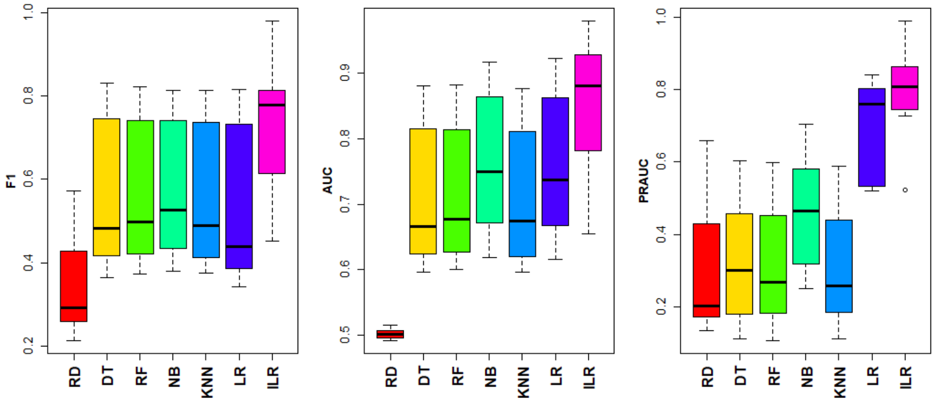


Fig. 3. Boxplots of F1 score, AUC and PRAUC on the next-4-semester dropout prediction after 2, 3, 4 semesters with different methods. The table shows ILR has a higher mean and lower variance across all metrics.

5 Experimental Results and Discussion

Fig. 4 shows the F1 and PRAUC for ILR approach and comparative baselines. In the cases of predicting next term, ILR has similar performance to standard LR. After adding the first estimated probability feature $\hat{\boldsymbol{p}}_3$, the F1 score improves by 34% and the PRAUC improves by 20% using only two semesters worth of data.

After normalization, the coefficient of $\hat{\boldsymbol{p}}_4$ is 9.25, which is more influential than the binary absence feature $\boldsymbol{A}_2$ (-2.1). Most other methods show poor results given data for just two semesters and ILR perform the best. ILR shows almost identical performance in F1 score when predicting 4th and 5th semesters dropouts with training sets of two and three semesters. Figure 4.4 is a boxplot

Table 4. F1 and PRAUC generated by prediction for cohort Fall 2013 students after 2, 3, 4 semesters with different methods. Table results shows out of 12 experiments, ILR wins 9 and 10 times for F1 and PRAUC respectively.

		F1 Score						
Given	Predict	RD	DT	RF	NB	KNN	LR	ILR
2	3	0.213	0.517	0.525	**0.544**	0.511	0.474	0.473
	4	0.226	0.449	0.47	0.51	0.465	0.405	**0.813**
	5	0.29	0.437	0.423	0.473	0.432	0.386	**0.778**
	6	0.296	0.42	0.442	0.44	0.413	0.374	**0.744**
3	4	0.241	**0.83**	0.822	0.813	0.813	0.815	0.814
	5	0.276	0.728	0.726	0.733	0.72	0.714	**0.778**
	6	0.295	0.692	0.698	0.697	0.691	0.677	**0.747**
	7	0.551	0.363	0.372	0.379	0.375	0.341	**0.485**
4	5	0.283	**0.791**	0.785	0.781	0.788	0.785	0.785
	6	0.306	0.763	0.757	0.747	0.755	0.75	**0.869**
	7	**0.556**	0.415	0.405	0.422	0.401	0.386	**0.452**
	8	0.572	0.411	0.418	0.428	0.41	0.392	**0.98**
Wins		0	1	2	3	0	0	**9**

		PRAUC						
Given	Predict	RD	DT	RF	NB	KNN	LR	ILR
2	3	0.135	0.289	0.262	0.312	0.261	**0.525**	0.524
	4	0.155	0.316	0.276	0.328	0.256	0.521	**0.726**
	5	0.2	0.32	0.319	0.366	0.295	0.528	**0.749**
	6	0.218	0.344	0.324	0.378	0.295	0.539	**0.741**
3	4	0.153	0.114	0.108	0.568	0.114	**0.84**	0.839
	5	0.189	0.198	0.193	0.556	0.188	0.747	**0.802**
	6	0.209	0.232	0.212	0.554	0.214	0.733	**0.886**
	7	0.633	0.604	0.598	0.705	0.589	0.77	**0.807**
4	5	0.197	0.13	0.146	0.25	0.146	**0.82**	**0.82**
	6	0.225	0.164	0.172	0.28	0.183	0.802	**0.888**
	7	0.633	0.572	0.58	0.596	0.583	0.786	**0.81**
	8	0.66	0.589	0.583	0.615	0.587	0.801	**0.989**
Wins		0	0	0	0	0	3	**10**

of F1, AUC and PRAUC. ILR perform the best in all three evaluation metrics with a higher mean and a smaller variance. From Fig. 2, we notice that the dropout rates are all below 25%. This is a highly imbalanced dataset. Precision and recall do not consider true negatives and thus won't be affected by the relative imbalance. Hence, both F1 Score and PRAUC are good at imbalance data on True Positive (TP), False Negative (FN) and False Positive (FP). PRAUC is more sensitive to False Positive than AUC. A low PRAUC prediction tends to identify a lot of students who are not going to dropout as target group.

We also report the counts of student across terms. Fig. 4 is a stack barplot in predicting dropouts given two semesters of test sets. False negative (black),

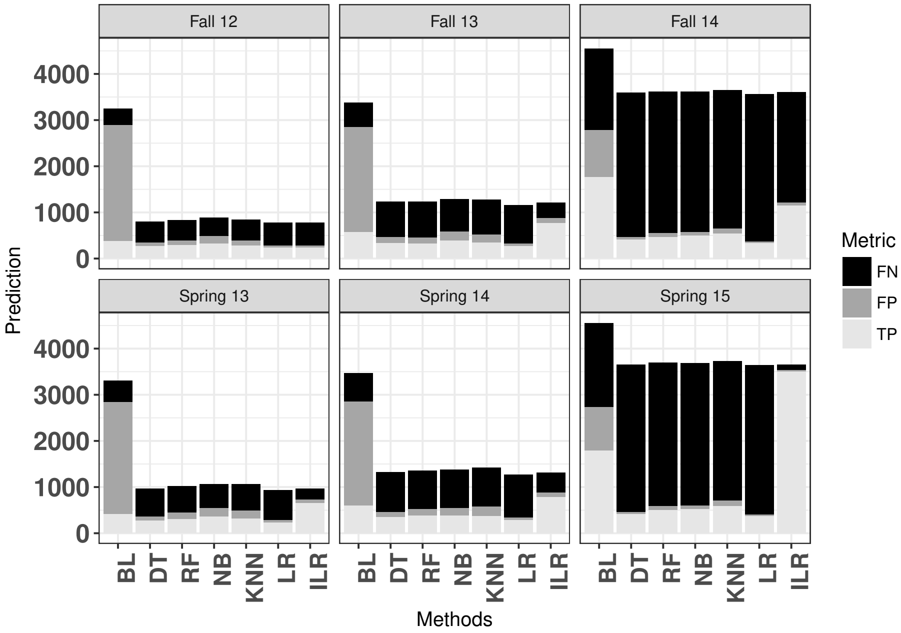


Fig. 4. Stacked barchart Given 2 terms dropout prediction of True Positive (TP), False Positive (FP) and False Negative (FN). For TP, the higher the better, FP and FN the lower the better.

false positive (dark grey) and true positive (light grey) correspond to the number of students we missed dropping out, successfully caught, and false alarms, respectively. The plots show that ILR has significantly greater area in correctly prediction of dropout (light grey) and less misses (black) while not producing many false alarm (dark grey). Therefore, we conclude that ILR performs significantly better in predicting future dropout.

5.1 Interpretation of coefficient

Since our model is essentially a logistic regression model, the coefficient of our model has direct interpretation. Standardized coefficients are usually useful for comparing the relative influence of different predictors within an logistic regression model [1]. To compute standardized coefficients, we divide raw values of the coefficients $\boldsymbol{\omega}_t$ by the standard deviation of their corresponding attributes. Table 5 shows the raw and standardized regression coefficient of ILR for predicting dropout in semester 6 after 2nd and 4th semesters. The former is an early identification and the latter is a late identification. We are interested in understanding how relevant our latent probability $\boldsymbol{p}_t$ compare to other given predictors such as GPA and absence in previous semesters. In the table, the coefficient of $\boldsymbol{p}_4$ and $\boldsymbol{p}_5$ are 17.0 and 4.87, and the magnitudes are much greater than the all other

Table 5. Raw and standardized coefficients of ILR for predicting dropout for 6th Semester with 2 (Early) and 4 (Late) semesters of data. ω_6^{Raw} is raw coefficient of ILR, where ω_6^{Norm} is standardized coefficient, which is not affected by scale of attributes. Attributes with '$*$' are generated by ILR.

Early prediction

Variables	ω_6^{Raw}	ω_6^{Norm}
(Intercept)	-1.236	.
$\boldsymbol{A}_2$	-0.248	-0.980
$\boldsymbol{G}_1$	-0.109	-0.107
$\boldsymbol{G}_2$	-0.034	-0.015
log(SAT_Total_1600)	.	.
HSGPA	-0.121	-0.14
log(ENTRY_AGE)	-0.0022	-0.0011
log(SAT_Math)	.	.
$\hat{\boldsymbol{p}}_3*$	.	. .
$\hat{\boldsymbol{p}}_4*$	**4.789**	**17.0**
$\hat{\boldsymbol{p}}_5*$	.	.

Late prediction

Variables	ω_6^{Raw}	ω_6^{Norm}
(Intercept)	-	-
$\boldsymbol{A}_2$	0.46	1.34
$\boldsymbol{A}_3$	0.34	0.94
$\boldsymbol{A}_4$	.	.
$\boldsymbol{G}_1$	-0.042	-0.0412
$\boldsymbol{G}_2$	.	.
$\boldsymbol{G}_3$	-0.102	-0.0766
$\boldsymbol{G}_4$	-0.0188	-0.144
$\boldsymbol{G}_5$	-0.677	-0.308
log(SAT_Total_1600)	.	.
HSGPA	.	.
log(ENTRY_AGE)	.	.
log(SAT_Math)	.	.
$\hat{\boldsymbol{p}}_5*$	**3.96**	**4.87**

attributes. This shows ILR generates dominant features from historical students records and successfully improved prediction greatly. That is why we have good performance of ILR in terms of F1 score, AUC and PRAUC.

5.2 Analysis on correlation

To analyze the effect of adding the previous latent probability p_{t-1} is added as a regressor, which we present in Fig. 5, the correlation plot. Stronger correlated variables are shown in dark color. Our model applies lasso on top of the features to remove highly correlated or irrelevant features to get a robust and interpretable model. From the plot, *SAT_MATH* and *SAT_1600* are strongly correlated and have not been selected twice in the coefficient table. From the top row of both plots, the colored grids of $\boldsymbol{p}_t$ against $\boldsymbol{y}_6$ shows that $\boldsymbol{p}_t$ are highly correlated with $\boldsymbol{y}_6$. We also observe that later semester has greater correlation with the label $\boldsymbol{y}_6$. The correlation plot agrees with our coefficient analysis and shows our features are relevant in the prediction.

6 Conclusion and Future Work

Predicting students at-risk is important for both, the institution and students. The event is rare and the timing is important. We propose an ILR model, learning from both the students' current semester-wise information as well as historical data from other students in the past with relatively small set of features. The

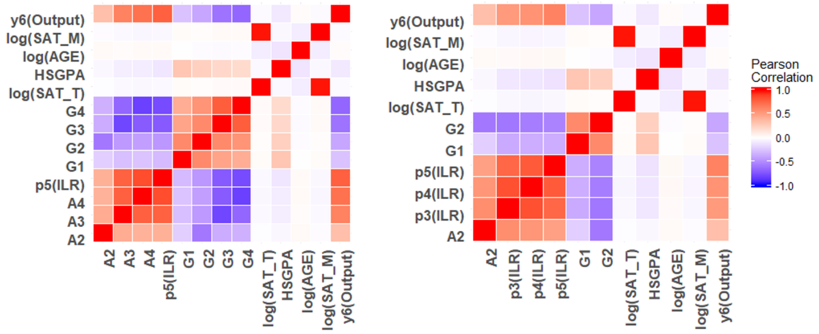


Fig. 5. Left Figure is the correlation between the regressors when predicting dropout in the 6-th semester given 2 semesters of data. Right Figure is the correlation plot for predicting dropout in the 6-th semester given 4 semesters of data (late). Darker color indicates higher correlation. $\hat{\boldsymbol{p}}_t$ are generated by ILR given t terms. $\boldsymbol{A}_t$ and $\boldsymbol{G}_t$ are known features of Absence and GPA for semester t. Result shows that the latent features $\hat{\boldsymbol{p}}_t$ have higher weights than given features both for early and late prediction.

coefficients of ILR can be normalized by dividing by standard deviation of the predictor variables to generate the variable importance for further interpretation. Our method has a few advantages compared to other methods.

First, it is an early prediction method which only requires a small amount of previous data from current students. Furthermore, because it uses the probability as a feature in the semester-wise trained model, we are able to take account of the previous "state" student performance and predict dropout for the next semester. The regularization of ILR features as well.

The proposed method will allow educational institutions to target student dropouts in a timely fashion and execute necessary actions accordingly. The model can be extended into coursework context with more available temporal information such as assignments, quiz and exams. This might lead to helpful interventions that help students and improve the overall educational quality and graduation rates.

7 Acknowledgement

This research work was supported by the National Science Foundation Grant 1447489. The experiments were running on ARGO, a research computing cluster provided by the office of research computing at George Mason University, VA (URL: `http://orc.gmu.edu`).

References

1. Agresti, A., Finlay, B.: Statistical models for the social sciences. Upper Saddle River, NJ: Prentice-Hall. Revascularization Procedures after Coronary Angiography. Journal of the American Medical Association 269, 2642–46 (1997)
2. Ameri, S., Fard, M.J., Chinnam, R.B., Reddy, C.K.: Survival analysis based framework for early prediction of student dropouts. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. pp. 903–912. ACM (2016)
3. Astin, A.W.: Predicting academic performance in college: Selectivity data for 2300 american colleges. (1971)
4. Baradwaj, B.K., Pal, S.: Mining educational data to analyze students' performance. arXiv preprint arXiv:1201.3417 (2012)
5. Cabrera, N.L., Miner, D.D., Milem, J.F.: Can a summer bridge program impact first-year persistence and performance?: A case study of the new start summer program. Research in Higher Education 54(5), 481–498 (2013)
6. Campbell, J.P., DeBlois, P.B., Oblinger, D.G.: Academic analytics: A new tool for a new era. EDUCAUSE review 42(4), 40 (2007)
7. Chen, Y., Johri, A., Rangwala, H.: Running out of stem: A comparative study across stem majors of college students at-risk of dropping out early. In: Proceedings of the 8th International Conference on Learning Analytics and Knowledge. pp. 270–279. LAK '18, ACM, New York, NY, USA (2018), http://doi.acm.org/10.1145/3170358.3170410
8. Druzdzel, M., Glymour, C.: What do college ranking data tell us about student retention? (1994)
9. Dynarski, M., Clarke, L., Cobb, B., Finn, J., Rumberger, R., Smink, J.: Dropout prevention. ies practice guide. ncee 2008-4025. National Center for Education Evaluation and Regional Assistance (2008)
10. Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. Journal of statistical software 33(1), 1 (2010)
11. Glynn, J.G., Sauer, P.L., Miller, T.E.: A logistic regression model for the enhancement of student retention: The identification of at-risk freshmen. International Business & Economics Research Journal (IBER) 1(8) (2011)
12. Golding, P., Donaldson, O.: Predicting academic performance. In: Frontiers in education conference, 36th Annual. pp. 21–26. IEEE (2006)
13. Hagedorn, L.S.: How to define retention. College student retention formula for student success pp. 90–105 (2005)
14. Horn, L., Carroll, C.D.: Stopouts or stayouts. Undergraduates who leave college in their first year (1999-087) (1998)
15. Kovacic, Z.: Predicting student success by mining enrolment data. (2012)
16. Lonn, S., Aguilar, S.J., Teasley, S.D.: Investigating student motivation in the context of a learning analytics intervention during a summer bridge program. Computers in Human Behavior 47, 90–97 (2015)
17. McFarland, J., Hussar, B., de Brey, C., Snyder, T., Wang, X., Wilkinson-Flicker, S., Gebrekristos, S., Zhang, J., Rathbun, A., Barmer, A., et al.: The condition of education 2017. nces 2017-144. National Center for Education Statistics (2017)
18. Nandeshwar, A., Menzies, T., Nelson, A.: Learning patterns of university student retention. Expert Systems with Applications 38(12), 14984–14996 (2011)
19. Pittman, K.: Comparison of data mining techniques used to predict student retention. Nova Southeastern University (2008)

20. Schneider, M.: Finishing the first lap: The cost of first year student attrition in america's four year colleges and universities. American Institutes for Research (2010)
21. Schneider, M., Yin, L.: The hidden costs of community colleges. American Institutes for Research (2011)
22. Seidman, A.: College student retention: Formula for student success. Greenwood Publishing Group (2005)
23. Stage, F.K.: University attrition: Lisrel with logistic regression for the persistence criterion. Research in Higher Education 29(4), 343–357 (1988)
24. Tanner, T., Toivonen, H.: Predicting and preventing student failure–using the k-nearest neighbour method to predict student performance in an online course environment. International Journal of Learning Technology 5(4), 356–377 (2010)
25. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) pp. 267–288 (1996)